DOCUMENT RESUME

ED 300 879 EA 020 233

AUTHOR Douglas, J. Rex; Johnson, M. Claradine
TITLE Interrater Reliability, Interrater Agreement, and Bias in a Nationally Sanctioned Assessment Center.
PUB DATE 86
NOTE 38p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Administrator Evaluation; *Administrator Selection; *Assessment Centers (Personnel); *Bias; Elementary Secondary Education; *Validity
IDENTIFIERS *National Association Secondary School Principals

ABSTRACT

Since 1975, the National Association of Secondary School Principals (NASSP) has attempted to improve the selection process for all levels of principals through the establishment of regional assessment centers. Presently, 37 centers operated by districts, consortia, universities, and state departments of education are producing data being used by approximately 4,000 school districts. The center being studied is university-operated, produces data for nine school districts, and is called the Center for the Assessment of Administrative skills (CAAS). This paper examines the internal validity of the process for selecting school principals at CAAS. NASSP supplied both the materials and the trained observers to be assessors. The study specifically examines (1) the variables of uniqueness and influence on overall ratings; (2) assessor reliability; (3) agreement of assessor ratings for each participant; and (4) persistent variance in participant performance for various demographic groupings. Results confirm that the eight centers conducted by CAAS have operated well within the standards required by NASSP, the sanctioning body, and within the broader standards required for effectively assessing candidates' potential. Further research is recommended to study assessor bias. Included are 17 references and 9 tables. (MLH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300879

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Rex Douglas

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Interrater Reliability, Interrater Agreement, and Bias in a Nationally Sanctioned Assessment Center

J. Rex Douglas, EdD

Associate Professor

Wichita State University

M. Claradine Johnson, PhD

Associate Professor

Wichita State University

BEST COPY AVAILABLE

EA 020 233

Abstract

The objectives of this study have been to examine the internal validity of an assessment center process for the selection of school principals as it has operated in one regional center, a sanctioned program of the National Association of Secondary Principals. Materials are supplied by NASSP, using NASSP-trained observers as assessors. Specific purposes of the study included (1) the examination of variables for uniqueness and influence on overall ratings, (2) the examination of interrater reliability, (3) the examination of interrater agreement, and (4) the examination of persistent variance in participant performance ratings for various demographic groupings. This study confirms that eight centers have been conducted well within the standards required by the sanctioning body, NASSP, as well as the broader standards required for an effective assessment of candidate potential.

Interrater Reliability, Interrater Agreement, and Bias in a Nationally Sanctioned Assessment Center

The National Association of Secondary School Principals (NASSP) has been engaged, since 1975, in an effort to improve the selection process for all levels of school principals through the establishment of regional assessment centers. There are now thirty-seven centers, producing data being used by approximately 4000 school districts. Centers are operated by districts, by consortia, by universities, and by state departments of education. The center being studied is university-operated, produces data for nine school districts, and is known as the Center for the Assessment of Administrative Skills (CAAS). Districts served include an urban district of 40000 students and eight medium-sized districts throughout the state. Eight centers have been conducted involving 48 assessor observers and 96 participant candidates.

Materials and procedures were developed and field-tested by NASSP personnel and their consultants before the first center was conducted in 1975. A three-year study by Neal Schmitt and associates (1982) clearly established criterion-related or predictive validity, content validity, and the internal validity of the process. Content validity and criterion-related

validity must be accepted when an agency assumes the responsibility of operating a regional center. Long-range plans include studies to re-confirm these aspects of the process, but the purpose of this effort has been to examine the operation of one regional center (CAAS) and deals with the various issues of internal validity.

Each assessment center has involved six assessors (trained by NASSP), twelve participants (candidates for the principalship) and six activities (two group activities, a fact-finding exercise, two in-basket exercises, and a structured interview). Each assessor observes two participants in each activity and prepares a definitive report of his/her behaviors. Two days are used for activities; three days are used to agree on the data and to assign ratings of one to five on each of twelve skill dimensions, an overall rating, and, in this center, a specific recommendation. The method of handling data is unlike virtually all assessment centers operating in other industries in that the time is spent in reaching agreement on exactly what the data says--reconstructing each event. Final ratings are then assigned by each assessor without consultation. The discussion, in other words, centers around the exact behaviors of the candidate in each activity, rather than a consensus discussion to decide on final ratings. This variation is extremely effective as a process, but final ratings remain as assigned unless distances between

ratings is too great. In the event that a two-point difference in ratings is assigned, assessors return to the data rather than negotiate consensus. Even though a one-point range is mandatory, differences can be more extensive than is likely when consensus discussions represent the final discussion. By an action of the center's board of directors, a participant must have acceptable ratings on three of five critical dimensions before he or she is recommended. These dimensions are Problem Analysis, Judgment, Organizational Ability, Leadership, and Sensitivity.

Four issues of internal validity were considered in the study:

1. Do the twelve dimensions each influence the overall rating of the candidate? Does each dimension represent a discreet set of information about the candidate?
2. Is there a high level of consistency, or reliability, between the ratings assigned by each of the six assessors for each skill dimension in each center conducted?
3. Is there a high level of agreement between the ratings assigned by each assessor for each dimension and for each participant in each center?
4. Are there persistent biases within each participant's ratings on skill dimensions

when participants are grouped by Sex, Race, Position, Degrees Held, Years of Experience, or Level?

Method

The Sample. The study of assessment centers has always been limited by the various pre-selection methods used as a basis for center participation. Participants in these eight centers were sometimes self-selected, sometimes district-selected, and, in all cases, district-sponsored. No assumptions can be made beyond the study population, and no generalizations from the data are appropriate. Since pre-selection methods at all centers are at least similar, it appears to be appropriate to make general, but not statistical, comparisons with earlier NASSP studies (Schmitt, 1981, 1982).

The Measures. Demographic variables in the study are clearly nominal data. Other variables are assigned ratings. The issue among measurement theorists continues as to whether such ratings represent ordinal or interval data. There is some agreement that consistent treatment is more important so that comparisons, at least, will be made within the same context (Guilford, 1965). The procedures related to the first issue treat the data as interval data (Pearson r's) to produce a factor-like matrix. To study interrater reliability and interrater agreement, the

data were treated as ordinal data (Spearman-Brown rho's and Kendall's W), the latter producing somewhat more stringent tests of agreement than other approaches (Ferguson, 1974). To examine bias, the data were treated as interval data for the convenience of more direct comparisons.

Bias. Bias is defined for this study as persistent variance, with no implication of pre-judgment. Each demographic circumstance involves different implications and those will be discussed in speculative terms. To study each area in terms of potential prejudice would require ex post facto designs to test every hypothetical possibility (Kerlinger, 1973), a considerable task that is beyond the scope of this paper. A problem of bias, and consequent prejudice, that should be resolved within the study involves concerns related to the assignment level of the participant. Some persons have voiced concerns about the impropriety of an elementary-background candidate being judged on materials developed for secondary candidates (Johnson & Douglas, 1985). The reality, of course, is that the original design of materials included the intention that problems pursued should be those common to all levels--elementary, middle or junior high, and senior high. Persistent variance based on the level of participants will be pursued if necessary.

Internal Validity. Kerlinger (1973) identified

internal validity as anything that affects the controls of a design. In particular, these include (1) maximization of experimental variance, (2) control of extraneous variables, and (3) the minimization of error variance. Because of the nature of the relationship between regional assessment centers and the sanctioning organization (NASSP), internal validity problems center on the operation of the center. External validity issues are important, of course, but can only be corrected in the design of materials and procedures, a national concern. The studies of Schmitt and others (1982) confirm the efficacy of the materials, procedures, and assessor training. The problems particular to the operation of a regional center focus on whether assessors continue to develop unique information, whether that information is reliable, whether assessors have a high level of agreement related to participant ratings, and whether persistent variance for differing groups represents accurate data or group pre-judgments.

Procedures. Eight assessment centers have been conducted by the Center for the Assessment of Administrative Skills involving 48 assessors and 96 participant candidates. Participants were asked to complete an information sheet at the beginning of each center. Ratings on twelve skill dimensions were developed by each set of six assessors for twelve

participants. Records were kept of the final rating designations by each assessor for each participant, together with each assessor's designation of the participant's overall rating as a potential candidate. Additionally, a record was kept of each assessor's designation of "Highly Recommended," "Recommended," or "Not Recommended." One data file was set up with participants as cases; another was arranged with the assessors as cases.

The first issue studied concerned the dual question, "Does each of the twelve dimensions represent discreet information, but yet influence the outcome of the overall rating?" This is a construct validity question in terms of the development of materials, but becomes an operations question once a center is conducted. To develop materials (and procedures), one could factor analyze all available data, eliminating those dimensions that do not "load" on the primary factor. After the first factor was identified as the primary or unifying factor, other dimensions would have been examined (with rotation) to determine their uniqueness as clusters of viable information. To examine the operation of this test procedure, an intercorrelation matrix was generated to attempt, in broad terms, to recreate a matrix similar to the original factor matrix used in development. Without the original to examine, Schmitt's work of a similar nature

(1982) was used for comparison.

The second issue--the question of consistency--was studied using SPSS procedures (Nie et al, 1983), Reliability. This program sequence was developed to produce Spearman-Brown correlations, on a split-half basis, for each of twelve skill dimension ratings for each center.

The question of interrater agreement has remained a gently seething argument for many years. Schmitt (1982) chose to use the alpha statistic as a measure of agreement. Kendall developed two measures specifically for that purpose, a Coefficient of Concordance (W) and a Coefficient of Consistence (K), both nonparametric methods based on ranked variables. Kendall's W seems to have remained as a reasonable measure, although some complain that it retains the influence of linearity and, thus, is too close to other correlation measures (Cicchetti et al, 1977, Robinson, 1960). Cicchetti has developed a method for continuous-ordinal and dichotomous-ordinal data (Cicchetti, 1977) that requires a specific data pattern fit and uses a system of weights. Robinson proposes the use of intraclass correlations using Pearson r's. Schmitt (1977) offered a promising possibility in posing agreement as a multivariate problem and explored a factor analysis approach with follow-up regression. Because of the nature of the sample (candidates only) and the nature of the data

(ordinal in most definitions), Kendall's W was selected as a reasonable approach to examine interrater agreement, with the data matrix rotated to treat assessors as cases. There does seem to be agreement among measurement theorists that this transformation is necessary. This change was accomplished manually, setting up a data base with assessors as cases identifying thirteen ratings for twelve candidates. Using the SPSS procedure (Nie et al, 1983), Npar Tests, Kendall's W's were calculated with significance measured by chi-square. This issue of agreement seemed best served by reporting on each center separately, since center operation is the primary task of the current study. W's were developed for each center based on six assessors, and twelve candidates' ratings on thirteen skill dimensions, with the first analysis of agreement by dimensions. The second set of Kendall's W's were developed to analyze agreement by candidates.

Collons (1968) stated that "...the smaller the level of significance, the more confident we can be that factors other than chance are causing the agreement among the rankings by the evaluators." With degrees of freedom quite high as in this data, the ratio-measure calculated as Kendall's W can be quite low, still producing levels of significance well below .05 and .01 levels. The analysis of data in the study included the identification of those elements for which levels of

significance failed to meet the traditional criteria of $p=<.05$, as well as identification of those elements not meeting the much more demanding level of significance of $p=<.001$. For the df levels of the two sets of Kendall's W calculations, the level of significance of $p=.001$ produces W-levels approximating .50. The W score of .50 was thus selected as an arbitrary, but more stringent, level of agreement for discussion.

To examine bias, the SPSS procedure (Nie et al, 1983), Oneway, was used, comparing mean scores on twelve dimensions plus the overall rating when participants were grouped by Sex, Race, Position, Degrees Held, Years of Experience, and Assignment Level. The program, Anova, was also run to examine two- and three-way comparisons. No interaction effects were indicated in any combination and these were not reported.

Results

Table 1 examines the intercorrelations of the twelve skill dimensions and overall rating for all of the 96 participants.

Insert Table 1 About Here

If we treat Rating as the primary or unifying factor, we would expect "loadings" to be generally above the .40 level for all variables. This was the case for all except the variables, Range of Interests (.24) and

Stress Tolerance (.37). A comparison with Schmitt's results (1982) shows .55 for Stress Tolerance and .35 for Range of Interests. Correspondence between the two sets was quite high, although Range of Interests would be considered low compared to Schmitt's data. Pearson r's between other variables were generally low to moderate on both sets of data, indicating that variables do represent unique information.

Table 2 is the representation of Spearman-Brown Split-Half Correlations (rho's) for each of eight centers for twelve dimensions plus overall ratings.

Insert Table 2 About Here

Seventy-four of the possible 104 measures were between .70 and 1.00. Four were below the .70 level, a number offering the chance possibility below .04. The variables involved were Educational Values (.43), Range of Interests (.66), and Decisiveness (.46 and .57), representing four different centers. In all other centers, and for other variables within these four, reliability coefficients were consistently above .70, with a majority above .90.

To examine interrater agreement, Kendall's W's were calculated for each assessor's ratings on each variable for twelve candidates for each center operated. Table 3 exhibits these results and provides measures of

agreement for each variable.

Insert Table 3 About Here

The purpose of examining agreement for variables is to study which variables are difficult to measure or which variables generate assessor differences based on pre-judged definitions. In four instances throughout Table 3, levels of significance failed to meet the traditional .05 significance test. All others (101 by count), range between p=.0256 and p=.00001. In fourteen instancess throughout Table 3, Kendall's W's fell below the .50 level (p=<.001), representing 13.4% of the possible cells. Ten scattered cells indicated W's between .50 and .60. Eighty cells indicated W levels between .60 and .97.

Policy in CAAS requires that five dimensions be considered as critical variables and receive first consideration. These are Problem Analysis, Judgment, Organizational Ability, Leadership, and Sensitivity. There were two instances (Center No. 5 and Center No. 7) in which one of these variables showed W's below .50. Below-fifty W's were shown for these variables: Problem Analysis (one center), Sensitivity (one center), Decisiveness (three centers), Range of Interests (two centers), Personal Motivation (two centers), Educational Values (four centers), and Written Communication (one

center). Overall Rating, the unifying variable, indicated no unacceptable W's, with one center's Rating W's below .79.

The examination of each candidate's thirteen ratings is more important to the study of agreement (and the operation of a center) than isolated variables. Table 4 is a compilation of Kendall's W's for each set of 6 assessors, 13 dimensions for 12 candidates.

Insert Table 4 About Here

Levels of significance for the W measure of agreement for all candidates ranged from $p=.0018$ $p=.00001$, well below the traditional tests of $p=<.05$ or $p=<.01$. Those W's below .50 were identified for discussion; however, it should be noted that significance levels of $p=.0018$ to $p=.0004$ indicate agreement that is totally acceptable, if not exemplary.

Agreement measured by Kendall's W's for ten participants fell below the .50 level ($p=.001$). Eighty-six candidates showed W's between .50 and .93. The general trend from the first to the eighth center showed improvement. There were two instances of below-fifty W's in Center No. 2 and three in Center No. 3. Centers No. 1, No. 4, No. 5, No. 7, and No. 8 involved one candidate each. Center No. 6 showed no candidates for whom there were W's below .50.

Tables 5 through 10 focus on persistent variance related to grouping differences. Table 5 illustrates Mean Ratings, Standard Deviations, and Significance of F Values for male/female groups.

Insert Table 5 About Here

There were significant differences in F Values, favoring females, on Overall Rating, Problem Analysis, Organizational Ability, Leadership, and Written Communication. Three of these are regarded as critical variables and Written Communication is clearly related to the outcome of many other variables (see Table 1).

Table 6 displays Means, Standard Deviations, and Significance of F Values for groupings based on race.

Insert Table 6 About Here

Significant F's were indicated on one variable only, Decisiveness. Blacks were rated lower on this variable than either Whites or Hispanics.

Table 7 shows Means, Standard Deviations, and Significance of F Values for groupings based on Position--teachers and non-teachers.

Insert Table 7 About Here

Significant F Values were indicated on the variables, Educational Values and Stress Tolerance. Scores on Sensitivity, favoring non-teachers, were approaching an agreeable level of significance, .069. One-way analysis of variance programs were run, but not here recorded, with non-teachers sorted as Assistant Principals, Counselors, and Central Office Specialists. No significant F values for differences between groups were shown.

Table 8 displays Means, Standard Deviations, and Significance of F Values for groupings based on Degrees Held.

Insert Table 8 About Here

There were no significant F values based on this grouping.

Table 9 shows Means and significant F values on all variables for groupings based on Experience. For want of space, Standard Deviations were not included.

Insert Table 9 About Here

There were no dimensions in which significant F values were computed. Computer runs, not here recorded, in which experience groups were collapsed also failed to produce significant F's.

Table 10 indicates Mean Ratings and Significant F Values for groupings based on the candidate's Assignment Level.

Insert Table 10 About Here

No significant F levels were produced. Some dimensions showed obviously higher ratings for Central Office personnel, but the variation was not significant. More cases would be desirable in studying this area further.

Discussion

Findings

1. An intercorrelation matrix of Pearson r's of the twelve NASSP dimensions and overall ratings for all candidates indicated that Rating was influenced by all dimensions, although r's for Range of Interests and Stress Tolerance were lower than some prefer. Intercorrelations of the twelve base dimensions showed moderate r's, an indicator of variable uniqueness. These findings were consistent with Schmitt's national study of NASSP assessment centers (1982).

2. Interrater reliability, based on Spearman-Brown rho's calculated separately for each center on a split-half basis, ranged above .70 on 100 of 104 possibilities. Four r's below .70 were found on Educational Values (.43), Range of Interests (.66), and Decisiveness (.46 and .57). These findings were also

consistent with Schmitt's study (1982).

3. Interrater agreement for thirteen variables, measured by Kendall's Coefficients of Concordance (W's), was acceptable in all cases for Overall Rating. For four of 104 possible cells, levels of significance were above the traditional .05 level. Considering the more stringent test--W=>.50; p=<.001--all dimensions were acceptable except for the following: Problem Analysis (one center), Sensitivity (one center), Decisiveness (three centers), Range of Interests (two centers), Personal Motivation (two centers), Educational Values (four centers), and Written Communication (one center). Only the first two of these are part of the five critical variables designated by center policy. The purpose of this portion of the study was to identify those variables that generate less agreement between assessors.

4. Interrater agreement for each candidate, measured on the basis of thirteen ratings assigned by assessors, and measured by Kendall's Coefficients of Concordance (W's), was acceptable by traditional tests (p=<.05) for all candidates. Considering the more demanding tests--W=>.50; p=<.001--W's nominally below the .50 level were indicated for Center No. 1 (one candidate), Center No. 2 (two), Center No. 3 (three), Center No. 4 (one), Center No. 5 (one), Center No. 6 (none), Center No. 7 (one), and Center No. 8 (one).

The operation of the most recent five centers involved lower W's for only four candidates, an important improvement.

5. Persistent variance, or bias, was found in performance ratings when candidates were grouped by Sex. Females had significantly higher ratings on Overall Rating, Problem Analysis, Organizational Ability, Leadership, and Written Communication.

6. Significant differences were found in one area, Decisiveness, when candidates were grouped by Race. Ratings on this variable were lower for Blacks than for either Whites or Hispanics.

7. When candidates were grouped by Position (Teachers or Non-teachers), there were significant F values, non-teachers having higher ratings, on Educational Values and Stress Tolerance.

8. There were no significant F values showing performance rating differences when candidates were grouped by Degrees Held.

9. There were no significant differences in performance ratings when candidates were grouped by Years of Experience.

10. There were no significant F values on performance ratings when candidates were grouped by Assignment Level (Elementary, Middle or Junior High, Senior High, Central Office).

Conclusions

1. The operation of the Center for the Assessment of Administrative SKills was found to be consistent with Schmitt's national study of NASSP assessment centers (1981) in terms of the uniqueness of variables, influence of all variables on overall ratings, and high levels of interrater reliability.

2. In terms of interrater agreement, there were four isolated incidents of low Coefficients of Concordance for specific variables. Both the data and experience suggest some difficulty in the measurement of particular variables, especially Decisiveness, Range of Interests, Personal Motivation, and Educational Values. It should be noted that low agreement levels were concentrated on limited dimensions and never on overall rating. For specific candidates, agreement levels were totally acceptable in all cases, although lowest W's were cited for possible improvement. Experience suggests that ten of ninety-six is not a high number of candidates for whom nominal disagreement might occur.

3. Significantly different performance ratings on five variables favored females over males. Three of these were in areas regarded as critical by this center. This finding is generally consistent with Schmitt's national study (1981), suggesting that there may well be a real difference between males and females in the quality of the pool of candidates available to schools or that the backlog of females now becoming candidates

includes some persons with extremely strong performance capabilities.

4. Only one dimension, Decisiveness, shows significant performance rating differences when candidates are grouped by race. This finding is quite different from Schmitt's study (1981) in which non-whites had poorer performance ratings in seven of the thirteen categories. This would suggest more effective pre-selection may be occurring by the districts in this center, that the operation of this center shows less residue of racial bias, or that the pool of Black candidates now includes persons with a higher level of skills.

5. There were no significantly different performance ratings for groupings based on Degrees Held. It should be noted that the relative position of these candidates does not offer a wide range of educational levels. Virtually all of these 96 candidates with Masters Degrees (the lowest level) have twenty or more hours beyond the degree. Since being at that level is a necessity to being a viable candidate, this study should not be considered as one defining the impact of educational level and performance potential.

6. Frequently stated objections to the assessment center approach suggests that those with higher positions or those with more experience in years have the advantage in participating in a center (Johnson &

Douglas, 1985). As in other studies of assessment (Schmitt, 1981), these fears were not supported in this data.

7. A persistent objection that is raised concerning the NASSP Assessment Center program has been that it is designed specifically for secondary participants and offers an unfair bias for elementary candidates. The design of materials includes a clear intention to avoid this problem. No significant F values were generated in these eight centers that would support such an objection. Other studies (Schmitt, 1981, and Johnson & Douglas, 1985) are in complete accord with this conclusion.

Recommendations

1. Further study is recommended of data collection, behavior identification, and variable definition for some of the twelve variables, especially Range of Interests, Personal Motivation, Educational Values, and Decisiveness. This should occur both locally and nationally.

2. When the variable, Overall Rating, is considered, there is no instance in which appropriate agreement standards were not met in the operation of these eight centers. The process would be improved with careful attention to those variables most difficult to measure--those listed above. The intercorrelation r's suggest that there may be an over-concentration on the

identified critical variables. The suggestions already made would improve that problem.

3. This study of the internal validity of the operation of the Center for the Assessment of Administrative Skills suggests that eight centers have been conducted well within the standards required by the sanctioning body, NASSP, as well as the broader standards required for an effective assessment of candidate potential. Continued study of internal validity, as well as studies of external validity, are recommended. Future research should continue and expand the study of bias, both in the terms identified here, and with specific reference to particular groups of assessors.

References

Bray, D. W. (1982). The assessment center and the study of lives. American Psychologist, 37, 2, 180-189.

Cicchetti, D. V., Aivano, S. L., & Vitale, J. (1977). Computer programs for assessing rater agreement and rater bias for qualitative data. Educational and Psychological Measurement, 37, 195-201.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. American Journal of Mental Deficiency, 86, 2, 127-137.

Collons, R. D. (1968). An Algorithm for the Coefficient of Concordance. Atlanta, Ga.: Georgia St. College.

Ferguson, G. A. (1959). Statistical Analysis in Psychology and Education. New York: McGraw-Hill Book Co., Inc.

Greenwood, J. M. & McNamara, W. J. (1967). Interrater reliability in situational tests. Journal of Applied Psychology, 51, 2, 101-106.

Guilford, J. P. (1965). Fundamental Statistics in Psychology and Education (4th Ed.). New York: McGraw-Hill Book Co.

Kerlinger, F. N. (1973). Foundations of Behavioral Research. New York: Holt, Rinehart and Winston, Inc.

Johnson, M. C. & Douglas, J. R. (1985). The impact of

principals' assessment centers on career opportunities for women. NASSP Bulletin, 69, 484, 105-111.

Lawlis, G. J. & Lu, E. (1972). Judgment of counseling process: reliability, agreement, and error. Psychological Bulletin, 78, 1, 17-20.

Nie, M. H., SPSS Inc. (1983). SPSSX Users Guide. New York: McGraw-Hill Book Co.

Robinson, W. S. (1960). The statistical measure of agreement. American Sociological Review, 1, 17-25.

Schmitt, N. (1977). Interrater agreement in dimensionality and combination of assessment center judgments. Journal of Applied Psychology, 62, 2, 171-176.

Schmitt, N., Meritt, R., Fitzgerald, M. P., & Noe, R. A. (1982). The nassp assessment center: a validity report. NASSP Bulletin, 9, 134-142.

Schmitt, N., Noe, R. A., Meritt, R., Fitzgerald, M. P., & Jorgenson, C. (1982). Criterion-Related and Content Validity of the NASSP Assessment Center. Reston, Va.: NASSP.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 22, 4, 358-370.

van der Kamp, L. J. T., & Mellenbergh, G. J. (1976). Agreement between raters. Educational and Psychological Measurement, 36, 311-317.

Table 1

Intercorrelation of Participants' Dimension Ratings

	Rating	Pa	Ju	Oa	Lead	Se
Problem Analysis	.77					
Judgment	.80	.72				
Org. Ability	.76	.70	.72			
Leadership	.75	.54	.60	.56		
Sensitivity	.59	.46	.43	.27	.50	
Decisiveness	.54	.44	.44	.68	.39	.21
Range of Interests	.24	.23	.13	.19	.24	.16
Personal Motivation	.43	.36	.37	.32	.51	.32
Educational Values	.49	.41	.46	.35	.44	.33
Stress Tolerance	.37	.31	.35	.17	.30	.35
Oral Communication	.65	.52	.55	.47	.67	.48
Written Communication	.67	.54	.59	.58	.56	.41

*$p > .05$; all others $p < .01$

Table 1 (Continued)

Intercorrelation of Participants' Dimension Ratings

	De	Ri	Pm	Ev	St	Oc
Range of Interests	.01*					
Personal Motivation	.21	.58				
Educational Values	.07*	.48	.51			
Stress Tolerance	.02*	.43	.51	.50		
Oral Communication	.34	.26	.58	.54	.51	
Written Comm.	.39	.26	.38	.38	.25	.60

*$p > .05$; all others $p < .01$.

Table 2

Interrater Reliability Coefficients of Assessment Skill Ratings, Spearman-Brown rho's, Split Half Basis. N=96

	Ctr. #1	Ctr. #2	Ctr. #3	Ctr. #4	Ctr. #5	Ctr. #6	Ctr. #7	Ctr. #8
Rating	.98	.99	.97	.97	.94	.95	.85	.96
Prob. Analysis	.93	.90	.95	.89	.96	.88	.78	.96
Judgment	.93	.94	.96	.93	.95	.88	.74	.96
Org. Ability	.84	.92	.98	.92	.92	.97	.92	.96
Leadership	.97	.94	.96	.97	.93	.98	.94	.88
Sensitivity	.97	.94	.89	.96	.90	.90	.93	.94
Decisiveness	.91	.92	.88	.93	.46	.72	.57	.84
Range of Int.	.94	.91	.66	.97	.92	.96	.99	.90
Per. Motivation	.84	.79	.84	.98	.88	.94	.75	.90
Ed. Values	.43	.89	.88	.99	.75	.98	.82	.90
Stress Tol.	.86	.96	.87	.95	.88	.99	.99	.90
Oral Comm.	.91	.96	.93	.99	.79	.91	.82	.91
Written Comm.	.92	.95	.94	.98	.93	.96	.97	.94

$p < .01$, all cases

Table 3

Kendall's Coefficients of Concordance (W), 6 Assessors, 12 Candidates, 13 Dimensions. Agreement by Dimension.

	Ctr. #1	Ctr. #2	Ctr. #3	Ctr. #4	Ctr. #5	Ctr. #6	Ctr. #7	Ctr. #8
Rating	.89	.97	.80	.85	.79	.79	.53	.81
Prob. Anal.	.76	.63	.75	.56	.79	.74	.48	.84
Judgment	.64	.74	.78	.74	.72	.77	.62	.74
Org. Ability	.70	.75	.89	.73	.66	.72	.62	.74
Leadership	.66	.78	.86	.88	.65	.88	.80	.65
Sensitivity	.77	.81	.73	.77	.43	.71	.78	.70
Decisiveness	.74	.57	.77	.83	.27*	.57	.40	.49
Range of Int.	.65	.62	.34	.89	.68	.48	.89	.73
Per. Motivation	.61	.47	.59	.99	.62	.62	.56	.22*
Ed. Values	.28*	.64	.42	.99	.36	.77	.72	.25*
Stress Tol.	.63	.77	.62	.67	.50	.99	.95	.50
Oral Comm.	.63	.73	.71.	.95	.53	.65	.62	.86
Written Comm.	.82	.83	.73	.85	.69	.85	.92	.46

*p=>.05 (Chi Square); all others, p<.001

Table 4

Kendall's Coefficients of Concordance (W), 6 Assessors, 12 Candidates, 13 Dimensions. Agreement by Candidate.

	Ctr. #1	Ctr. #2	Ctr. #3	Ctr. #4	Ctr. #5	Ctr. #6	Ctr. #7	Ctr. #8
Candidate 1	.69	.68	.58	.66	.65	.86	.76	.80
Candidate 2	.73	.82	.56	.93	.61	.58	.81	.62
Candidate 3	.76	.73	.45	.73	.65	.63	.68	.71
Candidate 4	.71	.43	.72	.84	.59	.68	.76	.93
Candidate 5	.72	.50	.71	.62	.68	.69	.64	.71
Candidate 6	.74	.51	.54	.71	.60	.60	.78	.53
Candidate 7	.84	.78	.67	.46	.65	.70	.85	.57
Candidate 8	.50	.56	.48	.83	.66	.75	.49	.75
Candidate 9	.74	.48	.73	.62	.63	.57	.80	.68
Candidate 10	.68	.65	.69	.92	.65	.68	.55	.75
Candidate 11	.44	.72	.48	.90	.59	.76	.62	.82
Candidate 12	.87	.86	.76	.72	.37	.58	.73	.36

Note. p=.0256 to .00001 (Chi Square) for all candidates.

Table 5

Means, Standard Deviations, and Significance of F Value of Skill Dimension Ratings by Candidate's Sex. N=96

	Male		Female		Significance
	Mean	SD	Mean	SD	of F
Overall Rating	2.905	.61	3.266	.50	.002*
Prob. Analysis	2.922	.62	3.175	.55	.037*
Judgment	3.019	.58	3.122	.44	.331
Org. Ability	3.209	.60	3.535	.49	.004*
Leadership	2.936	.74	3.419	.58	.001*
Sensitivity	3.052	.51	3.212	.51	.131
Decisiveness	3.568	.55	3.757	.47	.073
Range of Int.	3.276	.39	3.324	.44	.574
Per. Motivation	3.466	.48	3.533	.47	.491
Ed. Values	3.252	.45	3.201	.36	.541
Stress Tolerance	3.229	.47	3.544	.36	.804
Oral Comm.	3.356	.49	3.544	.53	.075
Written Comm.	2.952	.65	3.343	.53	.002*

*$p < .05$ or .01

Table 6

Means, Standard Deviations, and Significance of F of Skill Dimension Ratings by Candidates' Race. N=96

	White		Black		Hispanic		Signif.
	Mean	SD	Mean	SD	Mean	SD	of F
Rating	3.14	.56	2.93	.67	2.81	.67	.256
Prob. Ana.	3.12	.62	2.84	.43	2.77	.70	.164
Judgment	3.13	.51	2.98	.52	2.60	.53	.070
Org. Abil.	3.43	.56	3.27	.59	3.02	.60	.198
Leadership	3.24	.69	2.98	.66	3.18	.87	.428
Sensitivity	3.13	.52	3.15	.44	2.83	.49	.437
Decisiveness	3.75	.46	3.31	.47	3.50	.50	.006*
Range of Int	3.29	.41	3.31	.44	3.50	.47	.543
Per. Motiv.	3.49	.48	3.63	.47	3.43	.47	.551
Ed. Values	3.20	.40	3.38	.41	3.23	.52	.325
Stress Tol.	3.20	.42	3.35	.48	3.13	.14	.465
Oral Comm.	3.50	.49	3.36	.66	3.25	.52	.441
Writ. Comm.	3.20	.64	3.01	.55	2.88	.55	.345

*$p < .01$, all cases

Table 7

Means, Standard Deviations, and Significance of F of Skill Dimension Ratings by Candidates' Position. N=96

	Teachers		Non-Teachers		Signif.
	Mean	SD	Mean	SD	of F
Rating	3.017	.63	3.180	.51	.170
Problem Analysis	3.018	.64	3.094	.54	.532
Judgment	3.030	.60	3.116	.41	.443
Org. Ability	3.380	.59	3.377	.54	.979
Leadership	3.091	.76	3.298	.60	.145
Sensitivity	3.045	.54	3.237	.47	.069
Decisiveness	3.712	.52	3.615	.51	.362
Range of Int.	3.273	.43	3.332	.40	.491
Per. Motivation	3.436	.49	3.575	.44	.151
Ed. Values	3.130	.35	3.334	.43	.013*
Stress Tol.	3.124	.37	3.511	.44	.018*
Oral Comm.	3.402	.50	3.511	.54	.307
Written Comm.	3.157	.65	3.156	.58	.995

*$p < .01$

Table 8

Means, Standard Deviations, and Significance of F of Skill Dimension Ratings by Candidates' Degrees Held. N=96

	MA		EdS		PhD or EdD		Signif.
	Mean	SD	Mean	SD	Mean	SD	of F
Rating	3.04	.56	3.11	.64	3.39	.49	.301
Prob. Anal.	3.05	.63	3.01	.55	3.25	.49	.220
Judgment	3.05	.53	3.03	.56	3.27	.30	.597
Org. Ability	3.35	.59	3.34	.57	3.63	.36	.480
Leadership	3.10	.68	3.26	.73	3.47	.66	.391
Sensitivity	3.19	.53	3.01	.50	3.21	.48	.293
Decisiveness	3.62	.51	3.60	.58	4.01	.28	.111
Range of Int	3.26	.39	3.35	.42	3.44	.53	.408
Per. Motiv.	3.46	.48	3.49	.48	3.80	.31	.110
Ed. Values	3.20	.39	3.22	.41	3.39	.49	.491
Stress Tol.	3.18	.41	3.25	.41	3.36	.50	.517
Oral Comm.	3.44	.49	3.42	.51	3.64	.69	.521
Writ. Comm.	3.11	.65	3.14	.63	3.45	.35	.416

Note. $p > .05$, all cases

Table 9

Means and Significance of F of Skill Dimension Ratings by Candidates' Years of Experience. N=96

	0-5 Yrs. Means	6-10 Yrs. Means	11-15 Yrs. Means	16-20 Yrs. Means	Over 20 Yrs. Means	Signif. of F
Rating	3.42	2.96	3.07	3.26	3.04	.491
Prob. Anal.	3.42	2.93	3.13	3.14	2.79	.249
Judgment	3.33	2.89	3.10	3.14	3.07	.537
Org. Ability	4.00	3.22	3.40	3.46	3.30	.321
Leadership	3.58	3.03	3.22	3.31	3.05	.582
Sensitivity	3.00	3.10	3.18	3.24	2.90	.361
Decisiveness	4.00	3.64	3.67	3.63	3.71	.905
Range of Int.	3.00	3.32	3.29	3.40	3.20	.523
Per. Motiv.	3.42	3.55	3.44	3.60	3.48	.752
Ed. Values	3.25	3.12	3.22	3.84	3.16	.223
Stress Tol.	3.00	3.10	3.21	3.39	3.15	.197
Oral Comm.	3.58	3.23	3.50	3.62	3.32	.132
Writ. Comm.	3.33	2.97	3.18	3.27	3.12	.617

Note. $p > .05$, all cases

Table 10

Means and Significance of F of Skill Dimension Ratings by Candidates' Assignment Level. N=96

	Elem. Sch. Mean	Middle Sch. Mean	Senior High Mean	Central Office Mean	Sig. of F
Rating	3.05	3.12	3.05	3.53	.355
Prob. Analysis	3.08	3.00	3.00	3.47	.405
Judgment	2.96	3.14	3.08	3.47	.168
Org. Ability	3.35	3.31	3.38	3.90	.195
Leadership	3.26	3.18	3.04	3.60	.335
Sensitivity	3.09	3.16	3.14	3.14	.879
Decisiveness	3.64	3.59	3.74	4.07	.187
Range of Int.	3.29	3.36	3.27	3.20	.809
Per. Motiv.	3.48	3.58	3.49	3.23	.483
Ed. Values	3.18	3.34	3.22	2.97	.198
Stress Tol.	3.23	3.28	3.14	3.20	.642
Oral Comm.	3.45	3.62	3.29	3.43	.137
Writ. Comm.	3.14	3.22	3.06	3.47	.508

Note. $p > .05$, all cases